Representation Languages for Narrative Documents

Gian Piero Zarri
University of Paris IV/Sorbonne, France

INTRODUCTION

A big amount of important, "economically relevant" information, is buried into unstructured "narrative" information resources: This is true, for example, for most of the corporate knowledge documents (memos, policy statements, reports, minutes, etc.), for the news stories, the normative and legal texts, the medical records, many intelligence messages as well as for a huge fraction of the information stored on the Web. In these "narrative documents," or "narratives," the main part of the information content consists in the description of "events" that relate the real or intended behavior of some "actors" (characters, personages, etc.)—the term "event" is taken here in its more general meaning, also covering strictly related notions like fact, action, state, and situation. These actors try to attain a specific result, experience particular situations, manipulate some (concrete or abstract) materials, send or receive messages, buy, sell, deliver, and so forth. Note that in these narratives, the actors or personages are not necessarily human beings; we can have narrative documents concerning, for example, the vicissitudes in the journey of a nuclear submarine (the "actor," "subject," or "personage") or the various avatars in the life of a commercial product. Note also that even if a large amount of narrative documents concerns natural language (NL) texts, this is not necessarily true. A photo representing a situation that verbalized could be expressed as "Three nice girls are lying on the beach" is not of course an NL text, yet it is still a narrative document.

Because of the ubiquity of these "narrative" resources, being able to represent in a general, accurate, and effective way their semantic content—that is, their key "meaning"—is then both conceptually relevant and economically important: Narratives form, in fact, a huge underutilized component of organizational knowledge. This type of explicit yet unstructured knowledge can be, of course, indexed and searched in a variety of ways, but it requires, however, an approach for formal analysis and effective utilization that is neatly different from the "traditional" ones.

BACKGROUND

Usual ontologies—both in their "traditional" and "Semantic Web" versions (see the "Knowledge Representation" and "RDF and OWL" articles in this Encyclopedia)—are not very suitable for dealing with narratives. Basically, ontologies organize the "concepts"—that we can identify here with the important notions to be represented in a given application domain – into a hierarchical structure, able to supply an elementary form of definition of these concepts through their mutual generic/specific relationships ("IsA" links). A more detailed definition of the concepts is obtained by associating them with a set of binary relationships of the "property/value" type (e.g., a "frame"). The combination of these two representational principles is largely sufficient to provide a *static*, *a priori* definition of the concepts and of their properties.

Unfortunately, this is no more true when we consider the *dynamic behavior* of the concepts, that is, we want to describe their mutual relationships when they take part in some concrete action or situation ("events"). First of all, representing an event implies that the notion of "*role"* must be added to the traditional generic/specific and property/value representational principles. If we want to represent adequately a narrative fragment like "NMTV (an European media company) … will develop a lap top computer system…," besides asserting that NMTV_ is an instance of the concept *company_* and that we also must introduce an instance of a concept like *lap_top_pc*, we have to create a sort of "threefold" relationship; this relationship includes a "predicate" (like DEVELOP or PRODUCE), the two instances, and a third fundamental component, the "roles" (like SUBJECT or AGENT for NMTV_ and OBJECT or PATIENT for the new lap top system) used to specify the exact function of these two instances within the formal description of the event. Moreover, in an event context, we also must deal with those "*connectivity phenomena"* like causality, goal, indirect speech, co-ordination, and subordination that link together the basic "elementary events." It is very likely, in fact, that dealing with the sale of a company, the global information to represent is something like: "Company X has sold its subsidiary Y to Z *because* the profits of Y have fallen dangerously these last years *due to* a lack of investments," or, returning to our previous example, that "NMTV will develop a lap top computer system *to put* controlled circulation magazines out of business," or, that dealing with the relationships between companies in the biotechnology domain, "X made a milestone payment to Y *because* they decided to pursue an in vivo evaluation of the candidate compound identified by X." In computational linguistics terms, we are here in the domain of the "Discourse Analysis" which deals, in short, with the two following problems: (1) determining the nature of the information that in a sequence of statements goes beyond the simple addition of the information conveyed by a single statement; (2) determining the influence of the context in which a statement is used on the meaning of this individual statement or part of it.

It is now easy to imagine the awkward proliferation of binary relationships that sticking to the traditional ontological paradigm it would be necessary to introduce to approximate high-level notions like those of "role" and "connectivity phenomena."

Solutions for representing narratives in computer-usable ways that could move beyond a strict "binary" framework have, therefore, already been proposed in the past. In the context of his work—between the mid 1950s and mid 1960s—on the set up of a mechanical translation process based on the simulation of the thought processes of the translator, Silvio Ceccato (Ceccato, 1961, 1967), proposed a representation of narrative-like sentences as a network of triadic structures ("correlations") organized around specific "correlators" (a sort of roles). The correlators (100 or 200 in all, according to the different natural languages) included conjunctions and prepositions, punctuation marks, and syntactic/semantic relationships like subject-predicate, substance-accident, apposition, development-modality, comparison, and so forth. Ceccato also is credited to be one of the pioneers of the semantic network studies, even if the "official" beginning of this discipline is traditionally associated with the first publication, in 1966, of the Ross Quillian's thesis on "Semantic Memories" (Quillian, 1968). Basically, semantic networks are directed graphs (digraphs) where the nodes represent concepts, and the arcs represent different kinds of associative links, not only the "classical" IsA and property-value links, but also "ternary" relationships derived from Case Grammar in Linguistics (see Fillmore, 1966), and labeled as Actor, Object, Recipient, Instrument, and so forth. A panorama of the different conceptual solutions proposed in a semantic network context can be found in Lehmann (1992). In the 1970s, a sort of particularly popular semantic network approach was represented by the Conceptual Dependency theory of Roger Schank (1972). In this theory, the underlying meaning ("conceptualization") of narrative-like utterances is expressed as combinations of "semantic predicates" chosen from a set of 12 "primitive actions" (like INGEST, MOVE, ATRANS, the transfer of an abstract relationship like possession, ownership and control, PTRANS, physical transfer, etc.) plus states and changes of states, and seven role relationships ("conceptual case") in the Case Grammar style. Conceptual Graphs (CGs) is the representation system developed by Sowa (1984, 1999) and derived from Schank's work and other early work in the Semantic Networks domain. CGs make use of a graph-based notation for representing "concept-types" (organized into a type-hierarchy), "concepts" (which are instantiations of concept types) and "conceptual relations" that relate one concept to another. CGs can be used to represent narratives in a formal way, like "A pretty lady is dancing gracefully," and more complex, second-order constructions like contexts, wishes, and beliefs. CYC (see Lenat, Guha, Pittman, Pratt, & Shepherd, 1990) concerns one of the most controversial endeavors in the history of artificial intelligence. Started in the early 1980s as a MCC (Microelectronics and Computer Technology Corporation, TX) project, it ended about 15 years later with the set up of an enormous knowledge base containing about a million of hand-entered "logical assertions" including both simple statements of facts and rules about what conclusion can be inferred if certain statements of facts are satisfied. The "upper level" of the ontology that structures the CYC knowledge base is now freely accessible on the Web (*www.cyc.com/cyc/opencyc*). A detailed analysis of the origins, developments and motivations of CYC can be found in Bertino, Catania, and Zarri (2001, pp. 275-316).

NARRATIVE KNOWLEDGE REPRESENTATION LANGUAGE (NKRL)

With the exception of CYC and (very partially) of the Conceptual Graphs, the greater part of the solutions evoked in the last section concern mainly pure academic work, implying very sketchy forms of implementation. Narrative Knowledge Representation Language (NKRL) (Zarri, 1997, 2003) represents an up-to-date, fully implemented, and relatively complete solution to the problem of representing narratives without a too important loss of the original "meaning." NKRL innovates with respect to the usual ontology paradigm by associating with the traditional ontologies of concepts an "ontology of events," that is, a new sort of hierarchical organization